知识管理论坛 ISSN 2095-5472 CN11-6036/C
Knowledge Management Forum
E-mail: kmf@mail.las.ac.cn
http://www.kmf.ac.cn

【研究论文】

数字人文视角下多日记人物关系联合挖掘及可视化研究

——以西南联大相关日记为例

张锦胜 林泽斐
福建师范大学社会历史学院 福州 350117

摘要：[**目的 / 意义**] 联合挖掘与西南联大有关的多部名人日记，构建融合多部文献信息的西南联大社会网络图谱，以期通过多日记联合挖掘，发现更多的潜在社会关系，突破单日记社会网络挖掘的局限性。[**方法 / 过程**] 以 1938—1941 年间与西南联大相关的多部日记为语料，利用 Python 程序统计人物共现关系，使用 Gephi 构建多日记社交网络图谱。通过社会网络分析方法，对网络拓扑特征、人物中心度特征以及基于模块化和 K-core 的人物群体特征等进行分析和探讨。[**结果 / 结论**] 相较于独立日记挖掘，多日记社会网络联合挖掘显示出更明显的网络结构特征，更加去中心化，社会关系信息也更为丰富，可揭示出较为隐蔽的社交关系，在数字人文领域具有良好的应用价值。

关键词：数字人文；社会网络；文本挖掘；西南联大
分类号：G254

引用格式：张锦胜 , 林泽斐 . 数字人文视角下多日记人物关系联合挖掘及可视化研究 : 以西南联大相关日记为例 [J/OL]. 知识管理论坛 , 2022, 8(3): 171-182[引用日期]. http://www.kmf.ac.cn/p/342/.

名人日记作为一种历史文献，较为真实地记录了仅作者了解却不为大众所知的事务，并能够间接反映特定时期社会、政治、经济、文化等方面的背景信息，具有很高的史料价值。与传统的日记研究相比，数字人文视阈下的文本挖掘方法具有高效且直观的优势，能够从海量语料中快速构建人物社会网络，其中蕴含的社会关系信息可与其他史料相互印证，甚至得到新的发现。现有名人日记文本挖掘工作均基于独立日记开展。相较于单日记文本挖掘，多日记联合挖掘可以更好地揭示历史时期的社交网络结构和关键人物，有助于丰富历史人物研究的广度和深度。因此，多日记联合挖掘在名人日记文本挖掘研究中具有重要的价值。

西南联合大学（以下简称“西南联大”）是抗日战争打响后，我国重要的高等教育机构，

基金项目：本文系福建师范大学本科教改项目“新文科背景下的《数字人文》课程建设”（项目编号：I202101011）研究成果之一。
作者简介：张锦胜，硕士研究生；林泽斐，副教授，博士，通信作者，E-mail: linzf@fjnu.edu.cn。
收稿日期：2023-04-11 发表日期：2023-06-16 本文责任编辑：刘远颖



知识管理论坛
2023 年 第 3 期（总第 45 期）

曾培养了一大批优秀的思想家、科学家、文学家、实业家和社会活动家，被公认为中国高等教育历史上一颗璀璨夺目的明珠。西南联大的师生中不乏有记日记习惯的人士，这为深入探究西南联大的发展历程提供了可以相互佐证的参考。近年来，诸如《梅贻琦西南联大日记》《郑天挺西南联大日记》《西南联大求学日记》等相关日记相继出版，为构建融合多日记的西南联大社会网络提供了条件。

基于此，本研究对与西南联大相关的 4 部日记进行联合挖掘，首次通过多日记联合挖掘的方式构建面向历史研究的较大规模的人物社会网络，以人物关系为主要脉络，发现并提炼西南联大相关的多本日记中所蕴藏的知识，以期为数字人文视角下名人日记联合开发工作提供参考借鉴。

1 文献回顾

目前，面向数字人文的文本挖掘工作主要涉及以下几个领域：①作者归属与风格分析，相关研究多采用定量统计分析和计算机辅助技术，对作者在用词、句式等方面的特点进行分析，以此来鉴定作者身份和风格特征 [1-2]；②作品情感分析，相关研究利用自然语言处理技术和情感分析方法，挖掘文学作品中的情感特征，从而自动分析文学作品的情感倾向性 [3-5]；③社会网络分析与挖掘，相关研究多使用自然语言处理技术，从文学作品中抽取人物并构建社会网络，以此来研究文学作品中的人物关系和社会结构特征 [6-9]；④面向人文文献的基础自然语言处理（Natural Language Processing, NLP）任务研究，相关研究主要针对古籍资料等人文文献，利用传统机器学习方法和深度学习方法对词法分析 [10-12]、命名实体识别 [13-15] 等基础性 NLP 任务进行探索。

日记是一种私人记载形式，按照时间顺序记录了作者的亲身经历以及作者对人、事、物的看法，历来被认为具有直接史料的价值 [16]。传统日记研究主要涉及历史学、档案学、艺术学、军事学等多个学科领域。例如，R. F. Grattan 对英国陆军元帅阿兰布鲁克勋爵的战争日记运用比较方法，利用军事与管理理论得出一些关于如何提出成功战略的结论 [17]；张诗洋对《张彭春日记》进行了深入研究，通过对该日记的分析和解读，补充了中国早期话剧发展史以及张彭春本人戏剧思想的论据 [18]；吴景平则对《蒋介石日记》进行了详细研究，从而印证了国民党在抗战初期对日的态度 [19]。这些研究都以传统人文研究方法发掘了日记所承载的历史、文化和社会价值。

近年来，随着数字人文研究热度的不断攀升，文本挖掘和社会网络分析方法开始被应用于名人日记研究中，如 T. Cserpes 对 18 世纪匈牙利贵族 S. Károlyi 的日记文本进行社交网络分析，阐释匈牙利贵族的社交网络如何与这一时期出现的新型社会地位相联系 [20]；J. Zhou 等使用 LIWC 古汉语词典和 CC-LIWC 系统作为分析工具，量化分析曾国藩日记以探究其心理变化 [21]；宋雪雁、钟文敏对《王世杰日记》和《谭延闿日记》的文本挖掘，较为系统地对日记所蕴含的社交网络、地理位置、文本情感进行知识发现 [22-24]；黄紫荆等使用 BERT（Bidirectional Encoder Representation from Transformers）模型对《拉贝日记》进行情感极性识别，揭示了南京大屠杀前后拉贝的情感分布特征 [25]。

值得注意的是，目前针对名人日记的文本挖掘研究均基于独立日记开展，而单一日记承载的信息量相对有限。相比于单一日记，具有相似社会背景的多部名人日记具有更大的信息量，且可以相互印证，从而具有更高的挖掘价值。因此，本研究将采用多文本联合挖掘的方式，以西南联大师生日记中的人物关系作为挖掘对象，借助文本挖掘技术对西南联大师生的社会关系进行分析与可视化展示，以此对面向数字人文的多日记联合挖掘方法予以探讨。

2 西南联大日记社会网络构建

2.1 数据来源

西南联大是中国抗日战争后由北京大学、

知识管理论坛，2023(3): 171-182
DOI: 10.13266/j.issn.2095-5472.2023.015

清华大学、南开大学内迁设于昆明的一所综合性大学。自 1937 年 8 月建立到 1946 年 7 月 31 日停止办学，该校共存在了 8 年 11 个月。西南联大保存了抗战时期我国重要的科研力量，并培育了大量杰出的学生，其中不少人成为了世界一流的学者。

本研究以《梅贻琦西南联大日记》《郑天挺西南联大日记》《朱自清日记》《西南联大求学日记》4 部名人日记作为语料开展研究，各日记的基本信息见表 1。由于 4 部日记的起始年和终止年不尽相同，为控制时间的统一性，取各日记记载时间与 1938—1943 年的交集部分开展研究，这一时间跨度包含了西南联大 8 年办学时间中的 6 年，涉及日记文本共约 90 万字。

表 1 西南联大相关日记的基本信息

日记名称	作者	西南联大身份	记载时间	字数/万字
梅贻琦西南联大日记	梅贻琦	校务委员会主席	1941—1946年	21
郑天挺西南联大日记	郑天挺	总务长	1938—1946年	86
朱自清日记	朱自清	中文系教授	1924—1947年	40
西南联大求学日记	许渊冲	外语系学生	1938—1943年	39

4 部日记都是作者对个人生活的日常记录，具有鲜明的个人风格。其中，梅贻琦作为校务委员会主席，记录较为简洁；郑天挺先生作为教务长，记录的内容琐碎且细致；朱自清教授语言十分干练简白；许渊冲先生在学生时期更多地记录读书学习与日常生活，较为详尽。4 部日记分别以校长、总务长、教授、学生 4 个身份反映出西南联大从创立之初到逐渐发展的过程。

2.2 语料预处理

日记原文中对人物的记录有着许多姓氏、字号、职位、身份、昵称、学位等不同种类的省略及代称。针对这些省略及代称，本研究结合百科、日记注释、档案、历史文献等有关资料，通过对原文的研读，查找、校对资料，建立人物姓名与在日记中称谓的对照词表，示例见表 2，以该词表为基准，通过文本编辑器查找、替换功能将原文中的各种指代称谓替换为人物的正式姓名并逐一加以人工核对。

本研究使用基于 Python 的 NLP 工具包 PaddleNLP[26] 作为文本分词工具。为提高人名分词的准确性，通过设置自定义词典，将日记出现的所有人名存放于词典文件。根据分词处理后所产生的词性标签，去除其他无关的词汇，提取各句中带有实质意义的人名词汇。

表 2 西南联大日记人物称谓对照词表（部分）

梅贻琦西南联大日记		郑天挺西南联大日记		朱自清日记		西南联大求学日记	
原文指称	替换名称	原文指称	替换名称	原文指称	替换名称	原文指称	替换名称
莘田	罗常培	逵羽	樊际昌	隐、竹、妻	陈竹隐	家珍	赵家珍
伯苓	张伯苓	表哥	张耀曾	今甫	杨振声	兆凤	万兆凤
兆贤、蒋校长	蒋梦麟	友三	闻一多	无忌	柳无忌	匡南	刘匡南
龙主席	龙云	顾一樵	顾毓琇	公超	叶公超	绍祖	万绍祖
勉仲	查良钊	夫人	周俶	心恒、循正	邵循正	同端	林同端
陈部长	陈立夫	大夫	徐行敏	叔玉、萧君	萧蘧	叶教授	叶公超
矛尘	章廷谦	月涵	梅贻琦	霍先生	霍世休	茀生	涂茀生
郁文、妻	韩咏华	雪屏	陈雪屏	陈斠玄	陈中凡	适之先生	胡适



知识管理论坛
2023 年 第 3 期（总第 45 期）

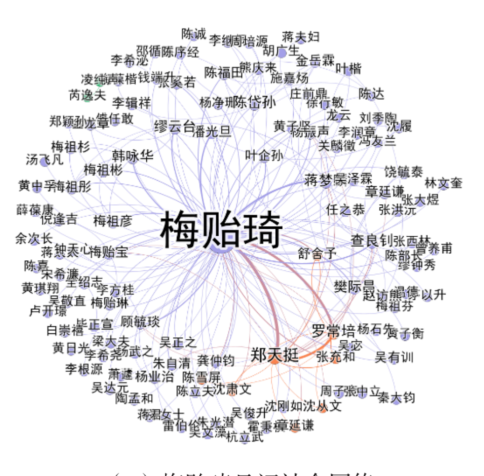
2.3 人物共现统计

为统计每一人名词汇对在所有句子中的共现频次，利用 Python 编程枚举每个句子中共现人名词汇对，然后将 4 部日记原文中所有句子中的相同人名词汇对进行归并统计。为了将分析重点聚焦于重要的高频人物，本研究通过阈值限定参与人物共现分析的人名数量，阈值设定为各日记及各年份出现频次最高的前 200 个人名词汇对中所出现的人名。

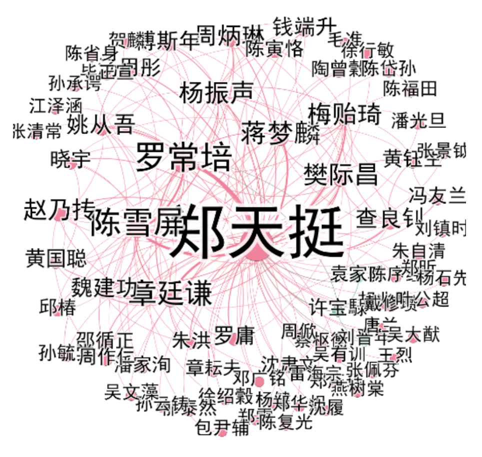
经过整理和统计，《梅贻琦西南联大日记》得到 118 名人物与 1 312 对共现关系；《郑天挺西南联大日记》得到 75 名人物与 6 718 对共现关系；《朱自清日记》得到 115 名人物与 1 040 对共现关系；《西南联大求学日记》得到 88 名人物与 1 568 对共现关系。四部日记综合去重后最终得到 317 名人物和他们之间的 10 638 对共现关系。最后，分别将 4 部作品及综合的人物共现关系转换为 CSV 格式的 Gephi 邻接表数据 [27]。

2.4 生成社会网络

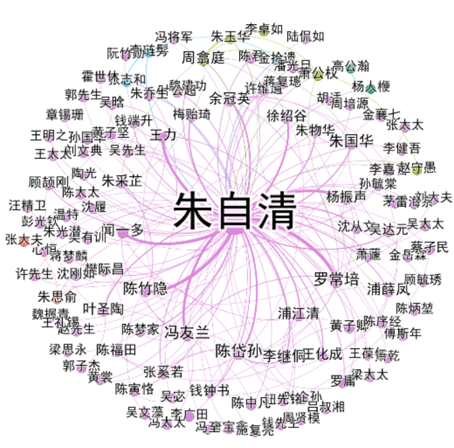
将处理过的 CSV 邻接表，导入 Gephi 中，使用 ForceAtlas 2 算法 [28] 生成各独立日记社会网络图谱以及融合各日记信息的社会网络图谱（以下简称“融合社会网络”），见图 1 和图 2。图谱中，节点的大小反映人物的中心度大小，边的粗细程度则反映出两个相关人物的共现频次。



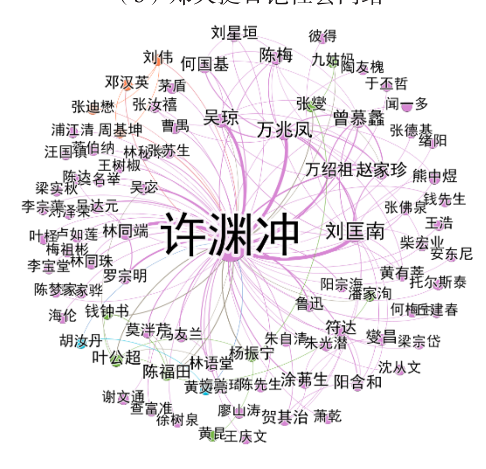
（a）梅贻琦日记社会网络



（b）郑天挺日记社会网络



（c）朱自清日记社会网络



（d）许渊冲日记社会网络

图 1 西南联大各独立日记的人物共现网络

ChinaXiv合作期刊
知识管理论坛，2023(3): 171-182
DOI: 10.13266/j.issn.2095-5472.2023.015

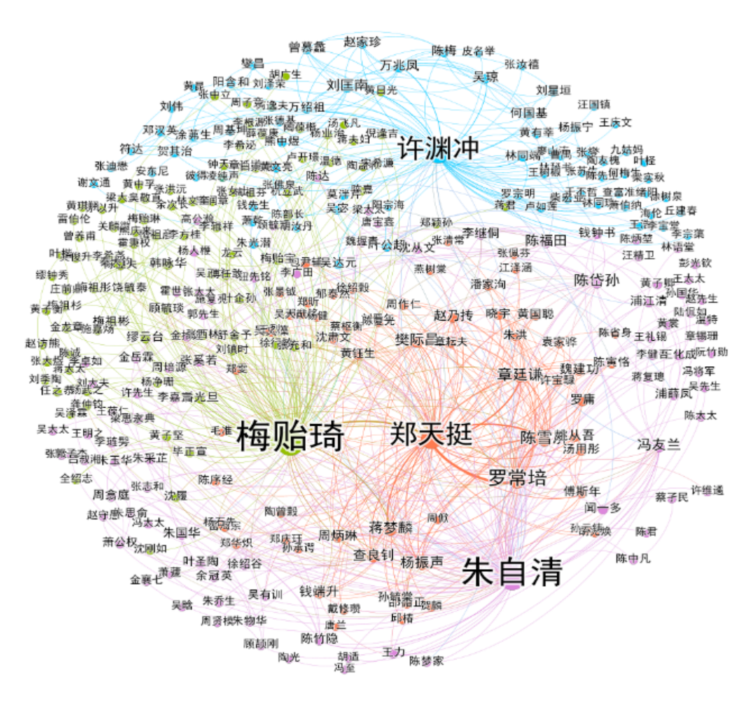


图 2 融合各日记信息的人物共现网络（融合社会网络）

本研究利用 Gephi 的网络统计功能，进行网络结构特征的计算。网络直径、网络平均度、平均路径长度等统计特征指标可以用来描述日记人物共现关系网络的拓扑结构。对本研究构建的多篇日记人物共现关系网络的特征指标进行计算，结果如表 3 所示：

表 3 西南联大日记人物共现关系网络特征指标

数据集	节点数	边数	网络直径	平均度	平均聚类系数	平均路径长度	图密度
梅贻琦日记	118	196	3	3.322	0.838	1.987	0.028
郑天挺日记	75	197	3	5.253	0.833	1.946	0.071
朱自清日记	115	190	4	3.304	0.708	2.101	0.029
许渊冲日记	88	157	4	3.568	0.762	2.066	0.041
融合社会网络	317	701	6	4.423	0.749	2.728	0.014

与单独日记社交网络相比，融合社会网络的规模较大，但是其网络密度有所下降，这与社会网络融合后人物数量增加，而 4 位日记作者的社交圈并非完全一致有关。根据网络统计特征指标计算结果可以看出，独立和融合社会网络的平均聚类系数均较大，平均路径长度较短，反映出典型的小世界特性 [29]，其中，融合社会网络的网络直径为 6，更接近真实世界社会网络的六度分隔（six degrees of separation）现象。

模块化指数（modularity index）为 M. E. J. Newman 和 M. Girvan 提出的社区划分评估指标 [30]。一般认为，模块化指数大于 0.3，即代表网络具有较明显的社区结构，真实世界社区的模块化指数通常介于 0.3—0.7 之间 [31]。本研究中利用 Gephi 划分社群并计算了融合社会网络的模块化指数，其指标为 0.352，这意味着同时进行多篇日记联合挖掘后，仍具有较为明显的网络社区结构。

根据生成的可视化图谱和网络结构特征，可以看出融合后的社会网络相较于单日记社会

网络更加地去中心化，不仅保留了各日记自身的社会关系，还揭示出一些较为隐蔽的社交关系，网络信息也更为丰富。

3 西南联大社会网络关系挖掘

3.1 网络人物中心度分析

3.1.1 融合社会网络人物中心度分析

本研究统计了融合社会网络中人物的中心度（degree centrality），排名前 20 位的人物见表 4。通过表 4 可知，4 部日记作者的中心度排名均为前列。郑天挺作为西南联大的总务长，既要负责外部关于西南联大的发展事务，也要负责学校内部教授的教学活动安排以及本身负责的教学领域研究，与各个教授、同仁交流较多；梅贻琦作为西南联大常务委员会的常委会主席，主抓西南联大的各项工作 [32]；朱自清原是清华大学中文系主任，在西南联大中与其他学校的中文系教授经常交流，且因学科建设研究要求等要与校长、总务长、文学院同仁保持联系；许渊冲是西南联大首届外语系学生，在社交网络中，他与同龄人、外文语文系的老师频繁交流。此外，还有樊际昌、蒋梦麟、杨振声、罗常培、罗庸、陈雪屏、章廷谦、姚从吾、查良钊等或是西南联大行政人员或是相关院系主要负责人，在人物共现图中占不小的比例。结合他们在西南联大的职务和身份，这一结果与西南联大历史事实相符。

表 4 融合社会网络中心度 TOP20 人物

人物	中心度	人物	中心度
罗常培	41	刘匡南	14
陈雪屏	25	赵乃抟	12
章廷谦	20	万兆凤	12
陈岱孙	20	周炳琳	11
蒋梦麟	19	钱端升	11
樊际昌	19	吴琼	11
杨振声	17	姚从吾	10
冯友兰	17	魏建功	10
查良钊	15	罗庸	9

3.1.2 各年份网络人物中心度变化分析

人物的社会关系在时间维度上具有一定的动态性，为分析不同年份西南联大的社会关系，我们融合各单独日记中的历年日记文本，生成各年份的融合社会网络。统计人物中心度信息后可以发现，各年份融合社会网络中核心人物中心度的变化，基本可以分为 3 类：①陈岱孙、查良钊、刘匡南、钱端升、吴琼、魏建功、曾慕蠡等核心人物在日记人物共现关系网络图中有着空白年份，即在该年份，由于该人物共现频次较低，未出现在所抽取的社会网络中；②樊际昌、杨振声、冯友兰、陈福田、赵乃抟、万兆凤、罗庸等核心人物在日记人物共现网络图中有着某一年份中心度与其他年份差异较大或各年份的中心度变化明显的现象；③罗常培、陈雪屏、章延谦、蒋梦麟等核心人物在日记人物共现关系网络中一直有着很高的中心度地位且变化幅度很小。

针对以上 3 种类型的人物，笔者从核心人物中选取部分有代表性的人物，代表性核心人物名单及其中心度相对排名的变化情况见表 5。

结合这些核心人物的生平及其在西南联大与梅贻琦、郑天挺、朱自清、许渊冲等日记作者或其他核心人物的交往过程进行分析，发现以上人物的中心度变化与历史领域学者对西南联大的许多研究成果相呼应。

在第一组中，刘匡南、吴琼、曾慕蠡受许渊冲参军入伍影响，中心度存在空白 [33]；陈岱孙是西南联大经济系的教授；查良钊、钱端升都是 1938 年应邀担任西南联大师范学院教授，查良钊次年出任联大训导长，钱端升之后出任北大办事处法学院院长，对应了查良钊与钱端升在人物共现网络中的中心度提升 [34]；魏建功与郑天挺原同为北京大学文学系教授，1940 年郑天挺兼任西南联大总务长事务繁忙，同年魏建功离职换岗，网络中心度变化体现了其二人事业方向的不同选择 [35]。

知识管理论坛，2023(3): 171-182
DOI: 10.13266/j.issn.2095-5472.2023.015

表 5 西南联大日记代表性人物中心度相对排名变化

人物		中心度相对排名					
组别	姓名	1938年	1939年	1940年	1941年	1942年	1943年
一	陈岱孙	9/137	57/211	——	26/285	32/251	14/251
	查良钊	——	78/211	13/149	33/285	4/251	5/251
	刘匡南	32/137	7/211	16/149	11/285	70/251	——
	钱端升	——	21/211	29/149	85/285	13/251	6/251
	吴琼	34/137	16/211	18/149	14/285	149/251	——
	魏建功	6/137	35/211	52/149	——	97/251	——
	曾慕蠡	——	39/211	23/149	12/285	——	——
二	樊际昌	4/137	9/211	3/149	9/285	18/251	82/251
	杨振声	12/137	6/211	1/149	80/285	5/251	2/251
	冯友兰	17/137	55/211	7/149	55/285	10/251	40/251
	陈福田	77/137	13/211	27/149	13/285	31/251	62/251
	赵乃抟	7/137	19/211	15/149	34/285	46/251	160/251
	万兆凤	58/137	12/211	19/149	4/285	214/251	232/251
	罗庸	8/137	20/211	63/149	69/285	37/251	102/251
三	罗常培	1/137	1/211	2/149	1/285	1/251	1/251
	陈雪屏	2/137	2/211	5/149	3/285	3/251	2/251
	章廷谦	5/137	5/211	8/149	7/285	2/251	8/251
	蒋梦麟	3/137	3/211	6/149	6/285	7/251	5/251

在第二组中，樊际昌在社交网络的中心度大幅下降是因为 1943 年国民政府令联大开办译员训练班，樊际昌担任训练班主任 [36]；杨振声在 1941 年中心度的下降印证了西南联大选派他任叙永分校校长的经历 [34]；冯友兰、陈福田均为许渊冲上过课，授课期间人物中心度大幅提升 [34]；赵乃抟的社交网络中心度在 1943 年骤降，印证了赵乃抟教授该年在译员训练班教学的经历 [37]；万兆凤与许渊冲联系紧密，直至 1942 年和 1943 年，许渊冲入伍，忙于毕业，其社交网络中心度迅速下跌；罗庸与郑天挺交流密切，至 1940 年郑天挺任教务长交流骤减，1942 年郑天挺将中文系事务交予罗庸，社交网络中心度的变化印证这一史实。

在第三组中，罗常培、陈雪屏、章延谦、蒋梦麟在社交网络的中心度常年稳定前列。罗常培是梅贻琦、郑天挺的左膀右臂，三人关系相当紧密，且罗常培与朱自清同为西南联大中文系教授 [34]；陈雪屏多次担任西南联大校务会议教授代表，曾任北京大学教育系代理主任，与郑天挺、罗常培等人交流频繁 [38]；章延谦曾任西南联大常务委员办公室秘书长，蒋梦麟历任中华民国教育部长、北京大学校长、西南联大常务委员会委员 [34]。

3.2 西南联大日记的网络人物社群分析

3.2.1 基于模块化的凝聚子群分解

调用 Gephi 统计设置中的 Community Detection 的模块化方法，选择节点—颜色—分割“Modularity Class”后不同社区有着不同的颜色，直观地验证了模块化后的结果：从网络整体出发，不同颜色之间的位置相对分明，有 4 个较为明显的社区，存在明显的人物社交群落，

见图 3。其中的左侧数字，为数据模块化后，分配给各个社区的默认 ID；右侧则是每个社区节点数占全部节点数的比例，从大到小，依次排列。

节点 边
统一的 分割 排名
Modularity Class

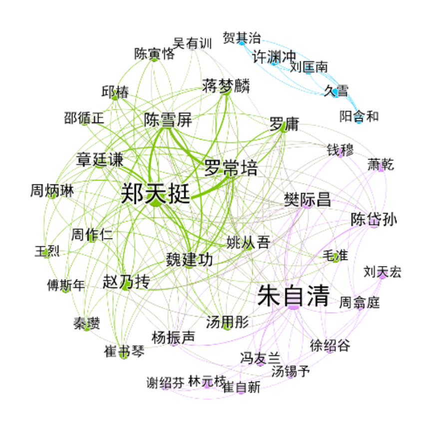
0	(28.75%)
2	(28.12%)
4	(22.36%)
1	(20.13%)
3	(0.64%)

图 3 凝聚子群分解结果

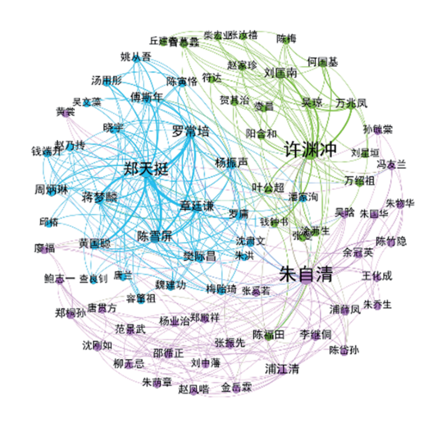
从融合社会网络图谱（图 2）可以看出，西南联大部分日记人物的共现关系有着鲜明的中心点，即 4 部日记的作者，因此该融合社交网络图谱有着明显的群体区分。此外，每个群体又有着更细致的分类，还具有明显的中心与外缘差别。所有人物都至少和 4 位中心人物有着关联，形成了复杂的社交网络。位于群体中心的人物彼此联系紧密，如：以梅贻琦、郑天挺为首的西南联大行政人员，掌握西南联大的外部事务及内部事务；以朱自清、胡适、罗常培为核心人员的西南联大中文系，进行频繁的学术交流和生活社交；以许渊冲、林同端、万兆凤、刘匡南为主要人员的西南联大学生，围绕学习和生活，占据了一部分人物共现图谱。而处于三大群体外的边缘人物只存在极少的人物共现关联，在西南联大影响力不足。

3.2.2 基于 K-core 的人物群体过滤

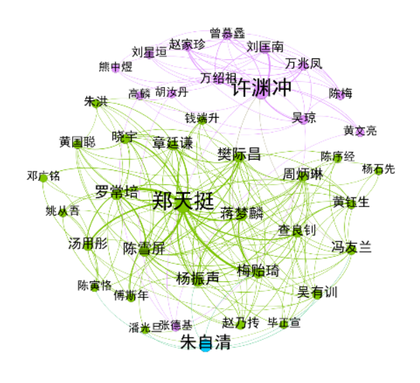
采用 K-core 对各年份模块化分解后的多篇日记社交网络进行过滤，可以更清晰地挖掘核心群体的人物及其之间的共现关系。在前文各个年份社交网络统计特征的基础上，分别以 K=3、4、5 进行观察，最后设置 K=4 为标准对 1938—1943 年各年份西南联大日记社交网络进行人物过滤，保留核心人物群体共现关系进行可视化展示，如图 4 所示：



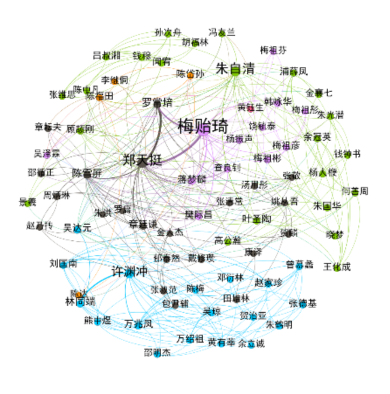
（a）1938 年



（b）1939 年

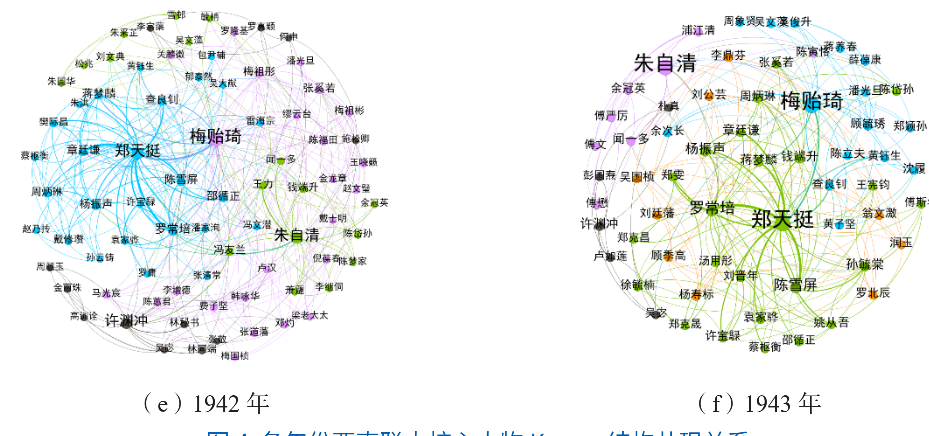


（c）1940 年



（d）1941 年

知识管理论坛，2023(3): 171-182
DOI: 10.13266/j.issn.2095-5472.2023.015



（e）1942 年　　（f）1943 年

图 4 各年份西南联大核心人物 K-core 结构共现关系

多篇日记人物 K-core 结构社交网络中，相同颜色节点代表人物是相同群体，不同颜色的节点位置距离程度代表人物群体间的关联密切程度。社交网络人物间的共现关系由边体现，链接权重越大，边越粗，意味着两个人物的共现关联越密切，联系越频繁。从整体趋势上来看，共现网络中人物节点受到力引导后能够较为清晰地划分为“学生”“教授”“行政人员”3类：“学生”部分人物多是与许渊冲相关联的人物或为江西籍学生，或为外语系学生，或为外语系教授，涵盖了许渊冲在西南联大求学的生活；“教授”部分人物或为西南联大同仁，或为朱自清的亲人朋友；“行政人员”或为西南联大常务委员、总务处、教务处和建设处等相关人员，或为政府行政人员。由上述分析可以看出，对不同年份的融合社会网络进行 K-core 过滤，能够较好地表现出各年份核心人物之间的社交关系变化。

3.2.3 核心人物群体分析

社会网络中的高中心度人物有较多的存留历史档案资料，通过分析这些资料，可以得出这些高中心度人物间的真实社会关系。本研究使用 Gephi 对各年份融合社会网络进行社区划分，得到各年份的人物类簇，最终选择群体较大的人物类簇进行分析，见表 6，并将各类簇中包含的核心人物（高中心度人物）与历史资料相印证。

表 6 中每年都在核心群体中占比较大的人物如蒋梦麟、章延谦、罗常培、魏建功、罗庸等人均是西南联大的教授或行政人员，与 4 位日记作者存在同事或师生关系 [34，39]。例如，1942 年群体 1 的缪云台是梅贻琦的好友，1938 年群体 0 中的邱椿、周作仁是郑天挺的好友，1938 年群体 1 的萧乾和 1941 年群体 1 的叶圣陶、闻宥、吕叔湘是朱自清的好友，1941 年群体 9 中的万绍祖、赵家珍、曾慕蠡、黄有莘、张德基是许渊冲的同学好友 [33，40-41]。1942 年群体 1 的韩咏华、梅祖彦、梅祖彬是梅贻琦的妻儿，1943 年群体 0 的郑雯是郑天挺的女儿，1938 年群体 1 的周翕庭是朱自清的姐夫，林同端是许渊冲当时爱慕的女生 [33]。1943 年群体 5 的陈立夫和 1942 年群体 1 费子坚、马光宸、张道藩、卢汉是国民政府下的军官或政府人员。从上述史料可以看出，各类簇中包含的核心人物在真实历史中具有明显的集聚性，这表明对本研究构建的融合社会网络进行社区聚类，可较为有效地反映出现实世界的人物社会关系。

此外，通过融合社会网络，还可以发现传统研究视角不易发现的隐蔽社群关系，例如结合图 2、表 6 和日记文本描述，可以较为直观地挖掘出西南联大桥牌社交网络：梅贻琦曾与缪云台、梅祖杉、萧蘧、章耘夫等 8 人打过桥牌；郑天挺曾与陈雪屏、罗常培、朱洪、周树人、章耘夫、邵循正等 12 人打过桥牌；朱自清曾与吴晗、萧蘧、陈岱孙、冯友兰等 19 人



打过桥牌；许渊冲曾与吴琼、陈福田等 25 人打过桥牌。不难发现桥牌这一娱乐活动在西南联大的师生中风靡一时，相关日记作者亦不例外：朱自清和吴晗、柳无忌、邵循正等十余人成立了桥牌俱乐部；许渊冲不仅平时和许多同学打桥牌，甚至和陈福田教授、联大几位助教都有过交手。与 4 位日记作者交手次数越多的人物，其在核心人物群体中的地位也越明显，例如，陈福田、陈岱孙、陈雪屏、陈省身被誉为西南联大桥牌名将“四陈”，1941 年群体 0 社群就包含了联大桥牌名将“四陈”中的两位。显然，小小的桥牌将西南联大的众多师生串联了起来，其在当时西南联大师生日常生活的重要性可见一斑。

表 6 各年份西南联大日记核心人物群体划分

年份	编号	核心人物	群体占比
1938	0	罗常培、陈雪屏、蒋梦麟、章廷谦、魏建功、赵乃抟、罗庸、姚从吾、周炳琳、汤用彤、邱椿、周作仁、陈寅恪、邵循正	14/46（35.1%）
	1	樊际昌、陈岱孙、杨振声、冯友兰、周翕庭、萧乾	6/54（41.2%）
1939	0	罗常培、陈雪屏、蒋梦麟、章廷谦、杨振声、傅斯年、樊际昌、周炳琳、黄国聪、晓宇、汤用彤、赵乃抟、罗庸、钱端升	14/53（25.2%）
1940	0	杨振声、罗常培、樊际昌、陈雪屏、蒋梦麟、冯友兰、章廷谦、周炳琳、吴有训、汤用彤、晓宇、查良钊、黄钰生、赵乃抟、黄国聪	15/52（35.1%）
1941	0	陈福田、陈岱孙	2/4（1.4%）
	1	叶圣陶、顾颉刚、朱国华、吕叔湘、余冠英、浦薛凤、杨人楩、钱穆、闻宥	9/70（24.6%）
	9	林同端、万兆凤、吴达元、刘匡南、曾慕蠡、吴琼、万绍祖、赵家珍、黄有莘、张德基	10/50（17.6%）
1942	0	罗常培、章廷谦、陈雪屏、查良钊、杨振声、邵循正、蒋梦麟、周炳琳、戴修瓒、樊际昌、许宝騄、雷海宗、冯文潜	13/56（22.3%）
	1	张奚若、梅祖彤、梅祖彬、缪云台、费子坚、韩咏华、卢汉、马光宸、张道藩	9/63（25.1%）
	7	冯友兰、王力、钱端升、闻一多、萧蘧、刘文典	6/58（23.1%）
1943	0	罗常培、陈雪屏、杨振声、孙毓棠、蒋梦麟、钱端升、章廷谦、姚从吾、周炳琳、袁家骅、许宝騄、陈岱孙、张奚若、刘晋年、郑雯、徐毓楠	16/63（25.1%）
	5	查良钊、顾毓琇、沈履、陈立夫	4/60（23.9%）

4 总结

本研究利用《梅贻琦西南联大日记》《郑天挺西南联大日记》《朱自清日记》《西南联大求学日记》等 4 部日记类书籍构建人物社会网络，从多本非结构化的日记文本中抽取出结构化人物实体与共现关联数据进行统计与量化分析，结合社会网络分析方法，对网络拓扑特征、人物中心性特征以及基于模块化和 K-core 的人物群体特征等问题进行分析与讨论，通过印证相关历史研究进行分析。与独立日记挖掘相比，多日记联合挖掘可以得出更明显的网络结构特征和更全面的社交网络可视化图谱，更加地去中心化，信息也更为丰富，有助于发现传统研究视角不易发现的隐蔽社交关系，从而对传统研究做出有益补充。

本研究也存在着一定的不足之处。本研究仅基于梅贻琦、郑天挺、朱自清、许渊冲 4 人的个人日记文本进行人物关系挖掘，师生关系、人员关系较为复杂，且该关系网络中的人物关系结构并不一定能够完全代表某一人物在当时西南联大师生群体间的影响力。梅贻琦的日记与其他 3 部日记存在时间不重合的问题，郑天挺和梅贻琦的日记原文存在一定的缺失，朱自清在日记中对人物共现情况的记录较为简略，

知识管理论坛，2023(3): 171-182
DOI: 10.13266/j.issn.2095-5472.2023.015

1943 年许渊冲在日记中的记录也特别简略，在一定程度上影响对人物共现原因的分析与判断。本研究所抽取的人物，多为西南联大文学院师生、西南联大行政人物，仅能展现西南联大局部师生关系、师师关系。此外，本研究所抽取的数据为局部时间段数据，仅能展现西南联大局部时间段之内的特定人物关系，更多、更丰富的人物关系的挖掘与呈现，还需更长时段的数据与更多相关日记文本的充实。

参考文献：

[1] 武晓春 , 黄萱菁 , 吴立德 . 基于语义分析的作者身份识别方法研究 [J]. 中文信息学报 , 2006(6): 61-68.

[2] 年洪东 , 陈小荷 , 王东波 . 现当代文学作品的作者身份识别研究 [J]. 计算机工程与应用 , 2010, 46(4): 226-229.

[3] LORD G, SMITH M N, KIRSCHENBUAM M G, et al. Exploring erotics in Emily Dickinson's correspondence with text mining and visual interfaces[C]// Proceedings of the 6th ACM/IEEE-CS joint conference on digital libraries. New York: ACM, 2006:141-150.

[4] 郎沁清 , 夏恩赏 , 饶高琦 , 等 . 数字人文视角下的金庸文本挖掘研究 [J]. 数字人文 , 2020(4): 115-136.

[5] Yumpu.com. Seeking the sentimental in nineteenth century American fiction[EB/OL]. [2023-03-12]. https://www.yumpu.com/en/document/view/33692161/seeking-the-sentimental-in-nineteenth-century-american-fiction.

[6] MORETTI F. Network theory, plot analysis[J]. New left review, 2011(68): 80-102.

[7] 范文洁 , 李忠凯 , 黄水清 . 基于社会网络分析的《左传》战争计量及可视化研究 [J]. 图书情报工作 , 2020, 64(6): 90-99.

[8] 宋雪雁 , 霍晓楠 , 刘寅鹏 , 等 . 数字人文视角下《全唐诗》贬谪诗人社会关系研究 [J]. 现代情报 , 2022, 42(2): 14-21.

[9] REYNALDO. Analyzing social networks of XML plays: exploring Shakespeare's genres - DH2018[EB/OL]. [2023-03-12]. https://dh2018.adho.org/en/analyzing-social-networks-of-xml-plays-exploring-shakespeares-genres/.

[10] 程宁 , 李斌 , 葛四嘉 , 等 . 基于 BiLSTM-CRF 的古汉语自动断句与词法分析一体化研究 [J]. 中文信息学报 , 2020, 34(4): 1-9.

[11] 程宁 . 基于深度学习的古籍文本断句与词法分析一体化处理技术研究 [D]. 南京 : 南京师范大学 ,2020.

[12] 李斌 , 袁义国 , 芦靖雅 , 等 . 第一届古代汉语分词和词性标注国际评测 [J]. 中文信息学报 , 2023, 37(3): 46-53.

[13] 于舒娟 , 毛新涛 , 张昀 , 等 . 基于词典和字形特征的中文命名实体识别 [J]. 中文信息学报 , 2023, 37(3): 112-122.

[14] 刘浏 . 古汉语典籍中的实体知识挖掘研究 [D]. 南京 : 南京大学 ,2018.

[15] 汤亚芬 . 先秦古汉语典籍中的人名自动识别研究 [J]. 现代图书情报技术 ,2013(S1):63-68.

[16] 齐世荣 . 谈日记的史料价值 [J]. 首都师范大学学报 (社会科学版), 2011(6): 1-15.

[17] GRATTAN R F. A study in comparative strategy using the Alanbrooke diaries[J]. Management decision, 2004, 42(8): 1024-1036.

[18] 张诗洋 . 新发现张彭春日记的文献价值考述 [J]. 文献 , 2021(5): 73-88.

[19] 吴景平 . 蒋介石与抗战初期国民党的对日和战态度 —— 以名人日记为中心的比较研究 [J]. 抗日战争研究 , 2010(2): 131-144.

[20] CSERPES T. Measuring identity change: analysing fragments from the diary of Sándor Károlyi with social-network analysis[J]. European review of history: revue européenne d'histoire, 2012, 19(5): 729-748.

[21] ZHOU J, ZHU T. Research on the psychology of historical figures based on big data analysis and data mining : taking Zeng Guofan's diary as an example[C]// Proceedings of 3rd international academic exchange conference on science and technology innovation. Guangzhou: IAECST, 2021: 704-708.

[22] 宋雪雁 , 崔浩男 , 梁颖 , 等 . 数字人文视角下名人日记资源知识发现研究 —— 以王世杰日记为例 [J]. 情报理论与实践 , 2021, 44(6): 105-111.

[23] 宋雪雁 , 钟文敏 . 数字人文视角下《谭延闿日记》人物关系挖掘及可视化研究 [J]. 情报科学 , 2022, 40(6): 25-35.

[24] 宋雪雁 , 钟文敏 . 数字人文视域下《谭延闿日记》的地理位置挖掘与可视化研究 [J]. 兰台世界 , 2021(10): 33-38.

[25] 黄紫荆 , 邱玉倩 , 沈彤 , 等 . 数字人文视角下的《拉贝日记》情感识别与分析 [J]. 图书馆论坛 ,2023,43(3):54-63.

[26] PaddleNLP Contributors. PaddleNLP: an easy-to-use and high performance NLP library[EB/OL]. [2023-03-01]. https://github.com/PaddlePaddle/PaddleNLP.

[27] Gephi. CSV Format[EB/OL]. [2023-03-02]. https://gephi.

org/users/supported-graph-formats/csv-format.
[28] JACOMY M, VENTURINI T, HEYMANN S, et al. ForceAtlas2, a continuous graph layout algorithm for handy network visualization designed for the Gephi software[J]. Plos one, 2014, 9(6): e98679.
[29] WATTS D J, STROGATZ S H. Collective Dynamics of ‘small-world’ networks[J]. Nature, 1998, 393(6684): 440-442.
[30] NEWMAN M E J, GIRVAN M. Finding and evaluating community structure in networks[J]. Physical review E, 2004, 69(2): 026113.
[31] GREEN D G, LIU J, ABBASS H A. Dual-phase evolution[M]. New York: Springer, 2014: 3-40.
[32] 于正阳 . 西南联大梅贻琦治校理念及实践研究 : 一个关系协调的视角 [J]. 扬州大学学报 (高教研究版), 2021, 25(3): 52-59.
[33] 许渊冲 . 西南联大求学日记 [M]. 北京 : 中译出版社 , 2021.
[34] 北京大学 . 国立西南联合大学史料：教职员卷 [M]. 昆明 : 云南教育出版社 , 1998.
[35] 杨绍军 . 魏建功先生在西南联大 [J]. 学术探索 , 2011(1): 2,145.
[36] 闻黎明 . 关于西南联合大学战时从军运动的考察 [J]. 抗日战争研究 , 2010(3): 5-18.
[37] 张友仁 . 赵迺抟教授的生平和学术 (下)[J]. 西安财经学院学报 , 2015, 28(2): 121-128.
[38] 刘火雄 . 兴观群怨 诗史互证——郑天挺西南联大时期的诗词交游及其学术活动考察 [J]. 文艺评论 , 2022(5): 17-25.
[39] 郑天挺 . 郑天挺西南联大日记 [M]. 北京 : 中华书局 , 2018.
[40] 吴卫萍 . 朱自清、叶圣陶的成都友谊 [J]. 青年文学家 , 2010(1): 24.
[41] 朱自清 . 朱自清日记 · 上 (1937-1941)[M]. 北京 : 石油工业出版社 , 2018.

作者贡献说明：

张锦胜：确定选题，提出研究思路，分析和处理数据，撰写论文，修改论文；

林泽斐：修改论文并定稿。

Joint Mining and Visualization of Character Relationships in Multiple Diaries from the Perspective of Digital Humanities——A Case Study of Diaries Related to Southwest Associated University

Zhang Jinsheng Lin Zefei

College of Social Development, Fujian Normal University, Fuzhou 350117

Abstract: [Purpose/Significance] By jointly mining multiple diaries related to National South-west Associated University (NSAU), a social network graph of NSAU that integrates information from multiple sources is constructed. The aim is to discover more potential social relationships through joint mining of multiple diaries, and break through the limitations of single diary social network mining. **[Method/Process]** Using multiple diaries related to NSAU from 1938 to 1941 as corpus, Python program is used to count co-occurrence relationships of characters, and Gephi is used to construct multi-diary social network graph. Through social network analysis methods, the network topology features, character centrality features and character group features based on modularity and K-core are analyzed and discussed. **[Result/Conclusion]** Compared with independent diary mining, multi-diary social network joint mining showed more obvious network structure features, more decentralized and rich social relationship information, which can reveal more hidden social relationships, and has good application value in the field of digital humanities.

Keywords: digital humanities   social network   text mining   National South-west Associated University